A brief report on protoDUNE online computing

Maxim Potekhin (BNL)

DUNE-LI Meeting

07/13/2016

Overview

- The ”protoDUNE Science Workshop” held at CERN on June 28th—30th was very useful (https://indico.fnal.gov/conferenceDisplay.py?confId=12042).
  - ...see next slides for details. Major impact on both online and offline.
- Current (but slightly out-of-date due to ongoing development) set of protoDUNE data parameters is kept as a spreadsheet in DUNE DocDB 1086, which is being updated based on the workshop results and some other follow-up.
- The “neut” cluster at CERN is now online and can be used if needed (initial configuration of 55 nodes with 300 more coming soon), access is rather straightforward.
- Renewed communications with the DAQ group about the DAQ interface to the online buffer.
- Experimented with a simple xrootd cluster setup (built from scratch) to understand what difficulties the DAQ group might face in implementing the interface.
- Discussions with the FNAL data experts to follow up on the data handling design such as documented in DocDB 1212.
- Received a request from the RACF leader Eric Lancon to provide estimates of the computing needs of protoDUNE@BNL for the next 5-10 years.

The “protoDUNE Science” Workshop

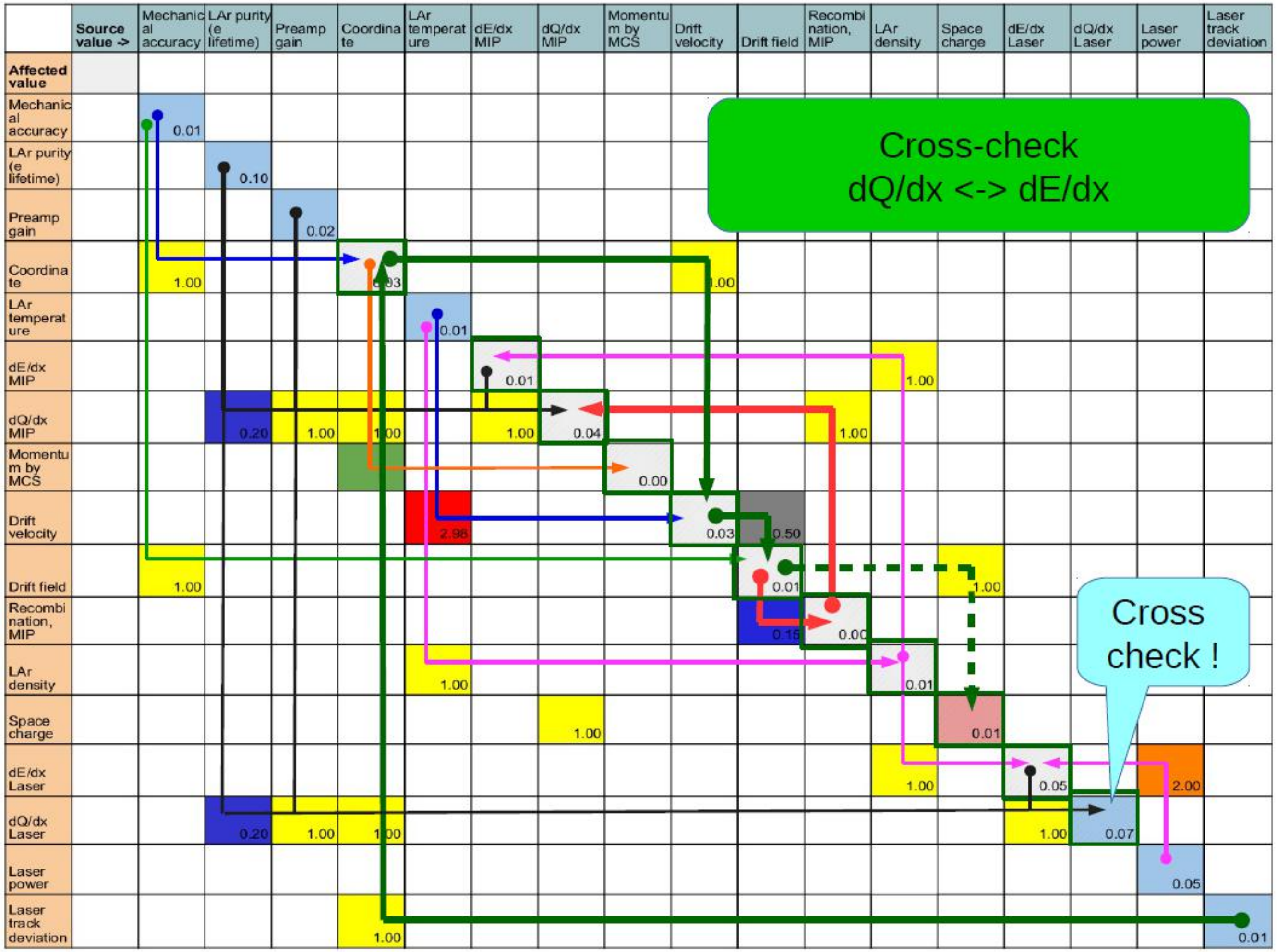
- Lots of good talks on physics including LARiAT
- a number of recent decisions announced (Andre/Thomas) such as:
  - the laser calibration system is no longer planned for protoDUNE due to cost and implementation risk considerations
  - no scintillator paddles for cosmic ray muons on top of the detector (cost and complexity)
  - muon trigger for particles close to the direction of the beam
  - scintillating fibers for tracking the beam particles (no MWPC)
  - total nominal number of triggers during the run upgraded to 25M to account for trigger inefficiency and other factors which of course are all still TBD
  - nominal trigger rate set as 25Hz in order to accomodate the higher statistics (rates as high as 50Hz are also discussed)
  - all 6 APAs are to be read out (different to a 3-APA readout proposed in Spring 2016)
- A coherent (but complex) calibration strategy was presented to reflect the new configuration sans laser. Implementation will require a major software effort.
- All of this not yet reflected in the TDR (ETA late summer) ― lots of updates will be needed.

Summary of impact on computing

- These updated protoDUNE parameters give a new basis for estimating the data characteristics and change the scale of the latter. This has a major effect on the scale and design of the online buffer and online/offline interface.and makes it more challenging. Similarly, the offline requirements are much higher (by an order of magnitude) — a comprehensive strategy is yet to be developed.
- Calibrations will require a well coordinated effort and manpower which as of yet has to materialize.

protoDUNE “Calibration Strategy with tracks”

- Please see the presentation by I.Kreslo at the workshop for details
- The screenshot below is meant to convey the complexity of the proposal



	Source value ->	Mechanical accuracy	LAr purity (e lifetime)	Preamp gain	Coordinate	LAr temperature	dE/dx MIP	dQ/dx MIP	Momentum by MCS	Drift velocity	Drift field	Recombination, MIP	LAr density	Space charge	dE/dx Laser	dQ/dx Laser	Laser power	Laser track deviation
Affected value																		
Mechanical accuracy		0.01																
LAr purity (e lifetime)			0.10															
Preamp gain				0.02														
Coordinate		1.00			0.03					1.00								
LAr temperature						0.01												
dE/dx MIP							0.01						1.00					
dQ/dx MIP			0.20	1.00	1.00		1.00	0.04				1.00						
Momentum by MCS									0.00									
Drift velocity						2.98				0.03	0.50							
Drift field		1.00									0.01			1.00				
Recombination, MIP											0.15	0.00						
LAr density						1.00							0.01					
Space charge								1.00						0.01				
dE/dx Laser													1.00		0.05		2.00	
dQ/dx Laser			0.20	1.00	1.00										1.00	0.07		
Laser power																	0.05	
Laser track deviation					1.00													0.01

Cross-check
dQ/dx <-> dE/dx

Cross check !

Impact of the “new configuration” on the online buffer

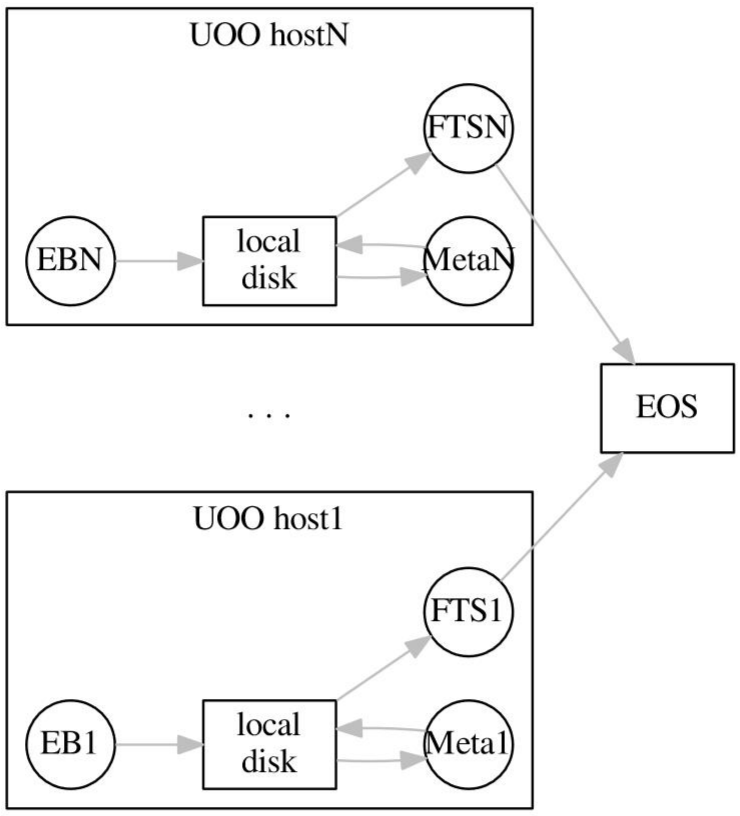
- Please see slides presented by Brett at a recent DAQ meeting: https://indico.fnal.gov/conferenceDisplay.py?confId=11233
- ...also spreadsheet in DocDB 1086. “Goldilocks scenario” officially dead.
- nominal lossless compression factor is assumed to be 4.
- trigger rates of 25—50Hz translate into instantaneous data rate of 1.5—3GB/s
  - can double when the final decision is made on the cosmic triggers
  - effective throughput of SATA III is 600MB/s
  - writing to multiple HDDs is unavoidable
  - throughput of NIC and switches is an additional consideration, cf. commodity switches and NICs are 1Gbps resulting in O(10) nodes necessary to have the required throughput
  - we are looking at up to 50 nodes to absorb the data

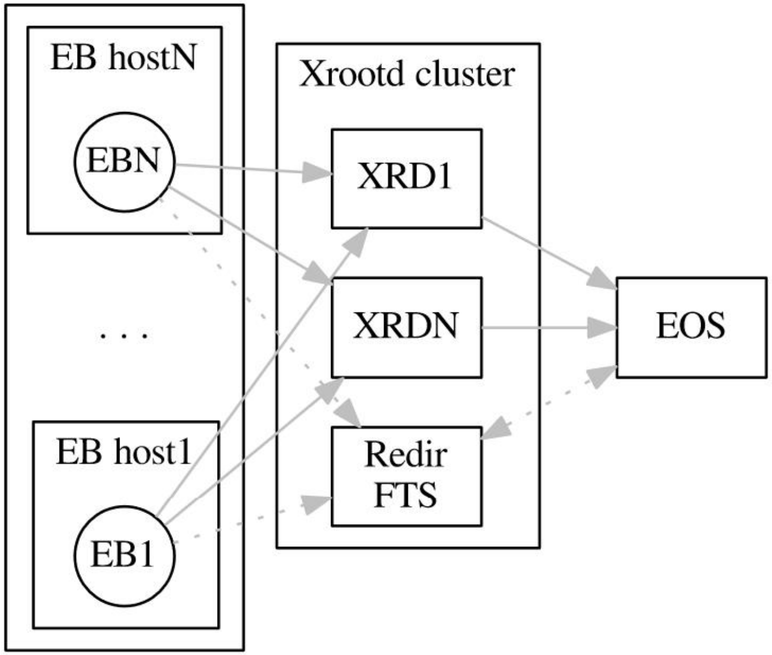
Design options for the online buffer

- two design approaches are currently being explored (Maxim & Brett):
  - "unified online/offline buffer" with HDDs attached to event builders and serving as input dropboxes for the FTS (file transfer system)
  - dedicated buffer networked to DAQ, with xrootd as primary candidate for clustering and load balancing: creates an additional layer in the data flow graph

- each has pros & cons:
  - **unified**: less hardware, more straightforward interface with online storage, but tight coupling between components with respect to design and planning/procurement cycle, less spare CPU
  - **dedicated**: more hardware and DAQ must have a more involved (but not too complicated) interface to storage; on the other hand decoupled design simplifies development (which can be done in parallel with FTS now), configuration and testing. Storage is decoupled from DAQ and is accessed essentially via a URI. Due to more nodes, there is more spare CPU available to do some sort of prompt processing and/or monitoring
    - IMHO enables us to do development in parallel and start earlier

Unified vs Dedicated online buffer





The online buffer action items

- A discrete event simulation of the DAQ/buffer/offline interface is in the works (Brett) which will allow us to understand bottlenecks, performance and scalability of various configurations under different assumptions.
- Working to understand the notification mechanism for the easiest integration of xrootd and FTS.
- Additional onus on DAQ to use xrootd is currently seen as minimal (only needing the client).
- Hardware options are being considered including repurposing a part of the “neut” cluster at CERN for xrootd, while upgrading the nodes with newer HDD, rough cost estimate $10k (just disk).
  - initial feedback from CERN (Nectarios) is favorable
  - for functional and some scalability testing the hardware can remain at its present location in the Idea Square building and/or in Bldg. 185
  - the plan is to continue development of xrootd configuration in a way that's best for interfacing both DAQ and FTS

Prompt processing

- We will likely want to closely monitor the noise characteristics and its evolution in time
  - noise spectra to be calculated continuously on a fraction of data (at CERN)
- Working event display is a must
- Right now unclear how much reco needs to be done in prompt processing mode
  - naively must do some, as one needs to monitor purity
  - coupling to space charge?

The “neut” cluster (recycled ATLAS TDAQ)

Impact of the "new configuration" on the protoDUNE offline

- Raw data (nominal as per DocDB 1086) is 1.5PB with beam triggers only
  - Compare with earlier estimate of 0.34PB in the "Goldilocks" scenario
  - considered by a few people in protoDUNE as the low limit of what will be taken
  - indications from FNAL that they will be ready to host a few PB worth of data (unclear how this will be supported)
- Calibrations group is still working on estimating the number of cosmic muon triggers that will be required, the best current estimate is about the same as the beam triggers
  - this doubles the amount and rate of data!

protoDUNE offline: resources

- AFAIK the protoDUNE request for FNAL resources was 8M CPU×hr/yr (TBD?)
  - a few reconstruction algorithms are under development now - not clear which one will end up in production and all are far from compelling optimization
  - ...so quite hard to estimate the necessary CPU requirement for time processing of the protoDUNE data - although the above request at least appears to be compatible with the "Goldilocks" scenario under some assumptions
- The "new normal" is up to an order of magnitude higher. What to do?
  - heard credible reports that OSG resources are already saturated
  - resources formerly available thorugh DOE allocations (like at PDSF) are likewise scarce or unavailable now
  - need to understand the limits (conservatively!) of what FNAL can provide
  - processing solely within the FNAL perimeter is unlikely now
- Resource federation and workload management may become an important requirement in order to leverage all that's available to DUNE
  - does not seem optional anymore due to a different scale
  - cf. promising contacts with BNL RACF and possible allocation here

Workload Management Systems (WMS)

- Recent initiative from FNAL SCD to evaluate WMS for protoDUNE
- A workshop is planned at FNAL on July 28th-29th to discuss this one more time and to meet with Panda experts in the same time frame.
- Nectarios (who built the "neut" cluster at CERN) is enthusiastic about unifying Grid resources for protoDUNE under a single management system.
  - there is feedback from the Czech group that there may be spare capacity that could be used by protoDUNE - proper WMS will make it a lot easier to "onboard" many users and production managers.
  - lxbatch and neut at CERN could be made available to users through the same interface and monitoring, along with BNL and perhaps the Czech cluster (also UK?)
- Investigating Panda
  - COMPASS at CERN have experience in running Panda on lxbatch (the public batch system).
  - Creating DUNE Panda queues at BNL should not be a problem due to considerable local expertise.
  - Consider this not the final technology choice but an evaluation exercise.

Offline Software

- Regardless of the reconstruction algorithms, operation principle of LArTPC dictates that signal deconvolution must be the first step in the reconstruction chain
  - has an effect on how and where this production step is done, preferably close to the data
  - after this step the volume of data will be greatly reduced since after deconvolution (including filtering) a threshold will be applied
- There is a to-do list covering software needed for the measurements program compiled by Robert and Dorota (see the workshop web page)
  - lots of work and effort... Volunteers needed
- Calibrations must be ready before the start of run and due to complexity appear to be a major challenge

BNL Computing: RACF

- We met with the new RACF director Eric Lancon and there is a spirit of cooperation
- RACF is willing to help with data handling expertise and computing support
- We are asked to formulate our requirements for the next 5 to 10 years, in terms of resources we need at BNL:
  - opinions?